UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions

Aciel Eshky[*1], *Manuel Sam Ribeiro*[*1], *Joanne Cleland*[2], *Korin Richmond*[1], *Zoe Roxburgh*[3], *James Scobbie*[3], *Alan Wrench*[3,4]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]Psychological Sciences and Health, University of Strathclyde, UK
[3]Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, UK
[4]Articulate Instruments Ltd., UK

{aeshky, sam.ribeiro, korin}@ed.ac.uk

Abstract

We introduce UltraSuite, a curated repository of ultrasound and acoustic data, collected from recordings of child speech therapy sessions. This release includes three data collections, one from typically developing children and two from children with speech sound disorders. In addition, it includes a set of annotations, some manual and some automatically produced, and software tools to process, transform and visualise the data.
Index Terms: Ultrasound and Acoustic Data, Child Speech, Disordered Speech, Speech Therapy.

1. Introduction

Speech sound disorders (SSDs) affect quality of life for a large number of children. In the UK, 11.4% of eight year olds[1] have persistent SSDs, ranging from common clinical distortions to speech that is unintelligible even to close family members [1]. SSDs are similarly prevalent in other countries [1]. Children with disordered speech experience adverse outcomes of many kinds: social and psychological outcomes, difficultly with literacy and educational attainment, and long-term employment prospects [2, 3, 4, 5].

However, current clinical practice for assessing these disorders is subjective and inaccurate [6]. Instrumental methods that use articulatory imaging, such as ultrasound, provide a more accurate diagnosis, at the expense of large amounts of manual effort from a highly trained pathologist. Machine learning has the potential to automate much of this work, leading to better outcomes for patients without increasing workload for pathologists, but publicly available data that could facilitate this work is scarce. Existing work reports results on adult data [7, 8, 9], data that is not publicly available [10], or data that is in proprietary format [11, 12, 13]. Additionally, child speech processing and disordered speech processing are both known to present many challenges [14, 15, 16, 17]. Having access to the right kind of data encourages more researchers to work on this problem and compare results.

In this paper we introduce UltraSuite, a curated repository of data obtained from child speech therapy sessions which used articulatory imaging techniques. The repository contains synchronised ultrasound and acoustic data recorded with a range of children with different categories of SSDs in addition to typically developing children who were learning new articulations. As part of the repository, we release annotations (manually and automatically produced) and software tools to process, transform, and visualise the data. The repository will continue to grow and become larger and more comprehensive as we add new studies, ensuring that all new data is available in the same standardised format. We invite other researchers to contribute their data to this repository.

1.1. Motivation

To better understand the potential for machine learning methods to automate the use of instrumental techniques for assessing, diagnosing and treating children with SSDs, it is useful to illustrate how instrumental techniques are used.

Perception-based methods for assessing SSD are known to be highly subjective [6]. Ultrasound imaging of the tongue provides additional information not available in the acoustic signal (e.g., the presence of double articulations or undifferentiated lingual gestures [18]). This additional information reduces subjectivity and in some cases changes the diagnosis. However, working with speech recordings and ultrasound videos is time consuming and difficult.

In order to provide a diagnosis or measure therapy progress, the speech pathologist goes through the following process: searching therapy recordings for occurrences of words of interest (the recordings often contain background noise, the voice of the therapist encouraging the child to speak, and the child uttering multiple words and potentially making mistakes); identifying boundaries of a phone of interest both in the audio and ultrasound video; locating the mid-frame of the phone in the ultrasound video; fitting a contour to the tongue in the image; and finally, measuring how similar/dissimilar the tongue shape is from the average tongue shape in correct articulation. While this process offers a more accurate diagnosis it is time consuming, tedious, and requires specialist training and is therefore not offered clinically. Automating it would allow it to be offered to children as standard practice.

1.2. Broader Applicability of the Data

Access to the articulatory domain through imaging techniques such as ultrasound gives additional information over the acoustic domain. Indeed, in addition to the area of speech and language pathology, prior work has shown that articulatory information has the potential to improve performance in multiple aspects of speech technology, for instance: speech recognition [19], speech synthesis [20], and silent speech interfaces [21].

[*]Equal contribution.
[1]Native speakers of English are expected to master production of all vowels and consonants by age 8 [1].

Table 1: *The number of participants, their gender and ages. We report ages in years (y) and months (m). We recorded the ages of participants on their single visit in UXTD, and on their first baseline in UXSSD and UPX.*

	UXTD	UXSSD	UPX
Number of participants	58	8	20
Female	31	2	4
Male	27	6	16
Mean age	9y 3m	7y 7m	8y 4m
SD age	1y 10m	1y 6m	2y 2m
Min age	5y 8m	5y 11m	6y 1m
Max age	12y 10m	10y 1m	13y 4m

1.3. Paper Outline

In Sections 2 and 3 we describe the process of collecting and standardising the data and include a description of a train-test split for a subset of the data. In Section 4 we describe our annotation work, followed by data statistics in Section 5, a description of the software tools in Section 6, and license information in section 7. We conclude in Section 8 with a brief description of data we are currently in the process of collecting with the aim of including in the repository in the future.

2. Data Collection

Ethical approval to collect the data was granted by the NHS Research Ethics Service. We recorded the data in the laboratory using the Articulate Assistant Advanced software (AAA), initially storing it in the AAA proprietary data format [22]. All sessions were conducted by a speech and language therapist (SLT), and both the children and the therapists spoke English with a standard Scottish accent. All therapists were female. We collected three datasets, one with typically developing children (TD) and two with children with speech sound disorders (SSD). The participants' guardians provided consent to allow the data to be made available to the research community[2].

A set of prompts specified the verbal task the child was expected to perform; for example, sentences to be read, isolated phones to be uttered, or pictures to be described. For each utterance, we recorded the acoustic signal and an ultrasound video of the child's mouth. We placed the ultrasound transducer probe submentally (under the chin) capturing the midsagittal view of the child's tongue [23] and stabilised it using a headset. Often the child needed encouragement to speak, so the acoustic signal contains both the SLT's speech and the child's speech. The child didn't always stick to the prompt; they hesitated, repeated, or made mistakes. In picture-describing tasks the speech is conversational in nature (e.g., "what's this?" "a frog stuck in a spiderweb" "aha, anything else?" "a strawberry in his mouth"). The prompt is therefore not a transcription of the audio.

2.1. Typically Developing Children

We recorded the typically developing subset of the Ultrax dataset (UXTD) between 11/2011–10/2012. The purpose of the experiment was to evaluate the effectiveness of ultrasound as a visual biofeedback tool for learning non-English articulations [24]. Each child attended once and recorded a single session.

[2]We excluded from the repository participants whose guardians did not provide consent.

2.2. Children with Speech Sound Disorders

The repository at this stage contains two datasets recorded with children with speech sound disorders. The first is the Ultrax speech sound disorders subset (UXSSD) which we recorded between 12/2011–07/2014, and the second is the UltraPhonix dataset (UPX), recorded between 06/2015–03/2017. The children exhibited a range of SSDs including phonological delay, phonological disorder, inconsistent phonological disorder, vowel disorder, articulation disorder, and childhood apraxia of speech. The data was recorded specifically for the purpose of evaluating the effectiveness of ultrasound as a visual biofeedback tool for therapy [25, 26]. Each child attended several sessions: suitability (before baseline), baseline (1–5 sessions), therapy (1–12 sessions), mid-therapy, post-therapy (immediately after therapy), and maintenance (several months after therapy). Table 1 shows the number of participants in each of the three datasets, their gender and ages. Persistent SSDs are more commonly associated with boys than girls [1], which explains the gender imbalance in the data.

3. Data Preparation

We exported the raw data from the proprietary AAA format to obtain a tuple of four files per utterance:

1. **Prompt file**: contains text describing the task the child was given and the date-time of recording.
2. **Audio file**: RIFF wave file, sampled at 22.05 KHz, containing the speech of the child and the SLT.
3. **Ultrasound file**: a sequence of ultrasound frames capturing the midsagittal view of the child's tongue. A single ultrasound frame is recorded as a 2D matrix where each column represents the ultrasound reflection intensities along a single scanline. The surface of the probe is convex and the scanlines are directed in an equal-angled fan in the scanning plane. In order to correctly interpret the ultrasound data, a set of parameters are recorded in the parameter file described below.
4. **Parameter file**: contains a set of parameters to interpret the ultrasound data and synchronise it with the audio. It gives the number of scanlines in each frame (63), the number of data points per scanline (412), number of bits used to represent each reflection intensity data point (8), the angle between each scanline (0.038°), the number of ultrasound frames per second (≈121.5 fps), and a synchronisation offset relative to the audio in seconds.

We discarded utterance tuples where the audio was too short and was unrelated to the prompt.

3.1. Prompts

We standardised the formatting of the prompt text by removing inconsistencies, such as replacing tabs with white spaces, removing duplicate or trailing white spaces, correcting the capitalisation of proper nouns, and correcting misspellings. We identified six distinct types of prompts:

(A) **Words**: a group of semantically unrelated English words (e.g., "down link pat get") which were either identified by the SLT as being diagnostically useful, or were based on a protocol from the Diagnostic Evaluation of Articulation and Phonology (DEAP) [27].

(B) **Non-words**: designed to elicit certain phones from the child but which are not real words (e.g., "p apa epe opo").

Table 2: *The number of utterances per prompt type, with the number of unique prompts in parentheses for each of the three datasets. We encode the type identifier in the file names.*

Type	ID	UXTD	UXSSD	UPX
Words	A	962 (26)	2708 (291)	3838 (455)
Non-words	B	607 (27)	495 (59)	560 (60)
Sentence	C	0 (0)	445 (35)	1020 (128)
Articulatory	D	2934 (45)	132 (17)	211 (31)
Non-speech	E	116 (2)	9 (1)	302 (1)
Other	F	0 (0)	56 (12)	61 (7)
Total		4619 (100)	3845 (415)	5992 (682)

(C) **Sentence**: designed to elicit co-articulation (e.g., "It's a toe Pam") or designed to examine phones of interest in different contexts at the sentence level (e.g.,"My Granny Maggie got a golden gown" where /g/ occurs in different word positions and different vowel environments).

(D) **Articulatory**: single or multiple phones occurring once or repeated. The SLT pronounces a phone, or plays a recording of a phone at different speeds, and the child is expected to imitate what they hear. The latter is known as a Diadochokinesis imitation task [28].

(E) **Non-speech**: includes swallowing motions recorded to obtain a trace of the hard palate, and coughs recorded to obtain additional tongue shapes.

(F) **Other**: conversational speech, such as describing a picture or telling a story about it (e.g., "Connected speech picture 1").

The number of utterances per prompt type in each dataset and the number of unique prompts are shown in Table 2.

3.2. File Naming Convention

We placed each session in a directory and labelled it accordingly (Suit, BL, Therapy, Mid, Post, and Maint). Typically developing children recorded a single session each, so the UXTD directories are labelled with speaker identifiers only. Within each session, we sorted the utterances by the date-time of recording and indexed them from 001. We then appended the prompt type identifier (A-F) to the index. For example, if the 5th utterance in a session is a sentence, the tuple of files associated with the utterances are 005C.txt, 005C.wav, 005C.ult, and 005C.param.

3.3. Train, Development, and Test Splits

We split the UXTD dataset into training, development, and testing subsets balancing by gender and age. Training contains 40 children (18 male, 22 female), development 6 children (3 male, 3 female), and testing 12 children (6 male, 6 female). The mean and standard deviation of the children's age in each subset is: 9y 5m $\pm$ 1y 10m, 9y 1m $\pm$ 1y 9m, and 8y 12m $\pm$ 1y 10m.

We omit a split for UXSSD and UPX due to the small number of participants in each SSD subcategory. However, we urge users of this data to report the ID of the participants and the name of sessions used for training and testing.

4. Data Annotation

In addition to the data described in the previous section, we release a set of annotations, including pronunciation dictionaries for each of the datasets, audio transcriptions for UXTD, SLT annotations, automatic speaker labelling and automatic phone alignments, all of which can aid modelling.

Pronunciation dictionaries: We prepared a pronunciation dictionary for each of the three datasets. We did this by listing the words that appear in the prompts, looking them up in a standard lexicon, and copying their phonetic transcription. We used a Scottish accent variant of the Combilex lexicon to match the accent in the data [29, 30]. For out-of-vocabulary words, such as the non-words of type B prompts, an annotator with training in phonetics transcribed their expected pronunciation. The vocabulary size is 296, 1048, and 1437 for the UXTD, UXSSD, and UPX datasets respectively.

SLT labelling: The SLTs annotated a small portion of the data for the intervention studies the data was originally collected for [24, 25, 26]. The annotations include boundaries of words and phones of interest, and tongue contours manually fitted to the mid-frame of phones of interest in the ultrasound video. The number of utterances with at least one label are 3900, 152, and 3919 in the UXTD, UXSSD, and UPX datasets, respectively. We release these annotation as Praat's TextGrid files [31] and follow the same naming convention described in Section 3.2.

Audio Transcriptions: Because the audio recording is not a direct match to the prompt, we provide a small subset of transcribed utterances, namely utterances of types A and B for all speakers in the UXTD dataset. A single annotator listened to the audio and transcribed the child's speech. SLT intervention is loosely transcribed as *[SLT:token]*, where *token* takes the form of *spn* (spoken sound) to denote generic SLT speech, or the form of a word if that word occurs in the prompt, for example *[SLT:helicopter]*. It is less obvious how to transcribe disordered speech, we therefore do not provide transcriptions for the UXSSD and UPX datasets.

Speaker labelling: In order to attribute different parts of the audio to different speakers, and to quantify the hours of speech, we trained a model that discriminates between SLT and child speech. We used the transcriptions of the UXTD dataset as training data by reducing words to *child* and *SLT* tokens, corresponding to turn-taking sequences between therapist and participant. Using Kaldi's [32] standard monophone recipe, we modelled these tokens with 5-state ergodic HMMs [33]. Silences were modelled with 5 state left-to-right skip HMMs. As a post-processing step, we merged identical labels that were separated by a silence with less than 100ms. A second pass of the data then removed labels with duration less than 50ms.

To measure the accuracy of this method, we used the force-aligned transcriptions of the UXTD's test set as a ground truth. We estimated error using *pyannote.metrics* [34], computing error in terms of seconds. We observed an Identification Error Rate of 4.6%, and precision and recall of 0.969 and 0.979, respectively. We decoded the three datasets with this method, which forms the basis for the data reported in Table 3.

Phone labelling: Automatically identifying phones in a child's speech would significantly reduce the workload for an SLT. As an initial solution, we applied standard phone alignment to the data. To obtain additional training data, we pooled the training subset of UXTD and the PF-STAR corpus [35], and trained a phone alignment model following the PF-STAR baseline recipe presented in [15].

Using our pronunciation dictionaries, we substituted the words in the prompts and audio transcriptions with their phonetic transcriptions for utterances of types A, B, and C. We then aligned the waveforms to the transcriptions in UXTD, and to

Table 3: *Hours of speech and silence rounded to two decimal places, estimated using the speaker labelling method described in Section 4, with percentages given in parentheses.*

	UXTD	**UXSSD**	**UPX**
Child speech	2.24 (28.39%)	3.66 (34.45%)	7.27 (38.70%)
SLT speech	1.24 (15.74%)	1.81 (16.99%)	1.92 (10.23%)
Total speech	3.47 (44.12%)	5.47 (51.43%)	9.19 (48.93%)
Initial silence	1.41 (17.96%)	0.91 (8.55%)	0.78 (4.17%)
Medial silence	1.99 (25.30%)	3.48 (32.69%)	7.11 (37.83%)
Final silence	0.99 (12.61%)	0.78 (7.32%)	1.70 (9.07%)
Total silence	4.40 (55.88%)	5.16 (48.57%)	9.59 (51.07%)
Total audio	7.87	10.63	18.78

the prompts in UXSSD and UPX since transcriptions are not available for these two datasets. Disordered speech is therefore aligned to expected pronunciation rather than true pronunciation.

Although we used standard methods to obtain phone alignments, these datasets pose significant challenges. Besides well known difficulties associated with child speech processing [14] and disordered speech processing [17], additional issues include variability in the recording conditions, interaction between the therapist and child, and deviations from the prompts. Future work will investigate more robust methods for phone alignment and ways of evaluating them.

5. Data Statistics

Overall, the data contains 37.28 hours of synchronised audio and raw ultrasound across all datasets. Table 3 shows the distribution of audio in terms of speech (child and SLT) and silences (utterance initial, medial, and final), estimated using the speaker labelling method described in Section 4. Although our speaker labelling method achieved good results on UXTD's test set, it was not directly evaluated on articulatory tasks or disordered speech. An inspection of the assigned labels in the data showed missed cases of child speech, especially when decoding utterances of type D. The estimates of speech shown in Table 3 are therefore conservative.

We estimated a total of 18.67 hours of speech in the three datasets. Despite this being a conservative estimate, it is comparable to the number of hours of speech in the standard child speech corpus PF-STAR, which has 10 hours of read speech by native English speaking children aged 6-11, and 10 additional hours of spontaneous speech by native English speaking children aged 4-14 [35]. We preserve initial and final silences in the data as the corresponding ultrasound may be useful for other tasks, such as tongue contour extraction. We estimated 91.83, 14.02, and 28.25 minutes of child speech for the training, development, and testing subsets of the UXTD dataset.

6. Companion Code Repository

We distribute a code repository containing a set of tools to interpret, transform and visualise the data, in addition to the recipes used to annotate the data. We describe the current contents of the code repository and invite users to contribute their own code.

Tools: The repository contains raw ultrasound reflection data, but we provide a set of tools to transform it for visualisation. A raw ultrasound file is a sequence of 2D matrices (or a 3D array) where each matrix is a frame, and each column in a frame contains ultrasound reflection data of a single scanline. To correctly interpret the ultrasound data, we provide a tool to transform the raw representation to the real world proportions. The function interpolates the spaces between the scanlines and the result is visualised as a fan image. Figure 1 illustrates this process. Another tool produces a sequence of images or a video from a raw ultrasound file.

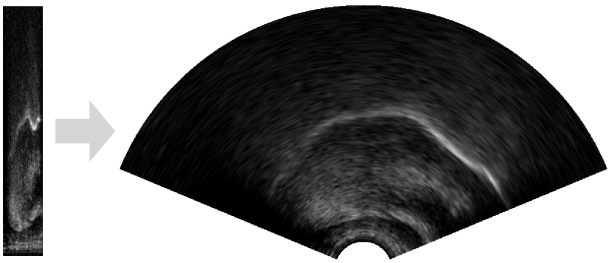

Figure 1: *An ultrasound image showing the midsagittal view of a child's tongue. We store the raw ultrasound reflection data efficiently as a matrix (left), but provide a tool to transform it to real world proportions (right).*

Recipes: We release the Kaldi recipes which we used to train the speaker and phone labelling models.

7. License and Distribution

We distribute UltraSuite under Attribution-NonCommercial 4.0 Generic (CC BY-NC 4.0) and distribute the companion code under Apache License v.2. Both can be obtained from the project website: `http://www.ultrax-speech.org/ultrasuite`

8. Conclusions and Future Work

We have introduced a new repository of ultrasound and acoustic data which we have collected from child speech therapy sessions. We have described the process of data collection, preparation and standardisation, along with a suggested train-test split. We have described tools to transform and visualise the data, and annotations including pronunciation dictionaries, audio transcriptions, SLT annotations, automatic speaker labelling and automatic phone alignments.

We will continue to grow the repository by adding more data and tools. We are in the process of collecting further data from 120 children with SSD in the Ultrax2020 project following the protocol described in [36]. In addition, we intend to add other available data to our repository, including adult data and alternative forms of articulatory imaging techniques (e.g., MRI of vocal tracts), all of which can be used in data augmentation methods [17, 15, 10]. We encourage other researchers to contribute by submitting their data for us to standardise and add to this repository.

9. Acknowledgements

Supported by: EPSRC Healthcare Partnerships Programme, grants number EP/I027696/1 (Ultrax) and EP/P02338X/1 (Ultrax2020), and NHS Scotland CSO, grant number ETM/402 (UltraPhonix). We thank Steve Renals for continued guidance and support, Anna Womack for help collecting the UltraPhonix data and Steve Cowen for technical support. We thank the children for participating and their guardians for providing consent.

10. References

[1] Y. Wren, L. L. Miller, T. J. Peters, A. Emond, and S. Roulstonef, "Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 4, pp. 647–673, 2016.

[2] A. Morgan, K. T. Eecen, A. Pezic, K. Brommeyer, C. Mei, P. Eadie, S. Reilly, and B. Dodd, "Who to refer for speech therapy at 4 years of age versus who to "watch and wait"?" *The Journal of pediatrics*, vol. 185, pp. 200–204, 2017.

[3] C. J. Johnson, J. H. Beitchman, and E. B. Brownlie, "Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes," *American Journal of Speech-Language Pathology*, vol. 19, no. 1, pp. 51–65, 2010.

[4] J. McCormack, L. J. Harrison, S. McLeod, and L. McAllister, "A nationally representative study of the association between communication impairment at 4–5 years and children's life activities at 7–9 years," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1328–1348, 2011.

[5] B. A. Lewis, A. A. Avrich, L. A. Freebairn, A. J. Hansen, L. E. Sucheston, I. Kuo, H. G. Taylor, S. K. Iyengar, and C. M. Stein, "Literacy outcomes of children with early childhood speech sound disorders: Impact of endophenotypes," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 6, pp. 1628–1643, 2011.

[6] S. Howard and A. Lohmander, *Cleft palate speech: assessment and intervention*. John Wiley & Sons, 2011.

[7] D. Fabre, T. Hueber, F. Bocquelet, and P. Badin, "Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] K. Xu, Y. Yang, M. Stone, A. Jaumard-Hakoun, C. Leboullenger, G. Dreyfus, P. Roussel, and B. Denby, "Robust contour tracking in ultrasound tongue image sequences," *Clinical linguistics & phonetics*, vol. 30, no. 3-5, pp. 313–327, 2016.

[9] D. Fabre, T. Hueber, L. Girin, X. Alameda-Pineda, and P. Badin, "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract," *Speech Communication*, vol. 93, pp. 63–75, 2017.

[10] D. Smith, A. Sneddon, L. Ward, A. Duenser, J. Freyne, D. Silvera-Tawil, and A. Morgan, "Improving child speech disorder assessment by incorporating out-of-domain adult speech," in *INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 2690–2694.

[11] N. Zharkova, "An ultrasound study of lingual coarticulation in children and adults," Dataset, 2009.

[12] ——, "High speed ultrasound/acoustic database of lingual articulation in preadolescents and adults," Dataset, 2011.

[13] ——, "High speed ultrasound/acoustic database of lingual articulation in typically developing children between three and thirteen years old," Dataset, 2016.

[14] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Workshop on Speech and Language Technology in Education*, 2007.

[15] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation." in *INTERSPEECH*, 2016, pp. 1598–1602.

[16] M. E. Beckman, A. R. Plummer, B. Munson, and P. F. Reidy, "Methods for eliciting, annotating, and analyzing databases for child speech development," *Computer Speech and Language*, vol. 45, pp. 278 – 299, 2017.

[17] H. Christensen, M. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech." in *INTERSPEECH*, 2013, pp. 3642–3645.

[18] F. E. Gibbon, "Undifferentiated lingual gestures in children with articulation/phonological disorders," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 2, pp. 382–397, 1999.

[19] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[20] K. Richmond, Z.-H. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview - application of articulatory movements using machine learning algorithms," *Acoustical Science and Technology*, vol. 36, no. 6, pp. 467–477, 2015.

[21] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[22] A. Wrench, *Articulate Assistant User Guide: Version 2.11*, Articulate Instruments Ltd., QMU, Musselburgh, United Kingdom, 2010.

[23] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics and Phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.

[24] J. Cleland, J. Scobbie, S. Naki, and A. Wrench, "Helping children learn non-native articulations: the implications for ultrasound-based clinical intervention." in *Proceedings of the 18th International Congress of Phonetic Sciences. ICPhS 2015*, 2015, pp. 1–5.

[25] J. Cleland, J. M. Scobbie, and A. A. Wrench, "Using ultrasound visual biofeedback to treat persistent primary speech sound disorders," *Clinical Linguistics and Phonetics*, vol. 29, no. 8-10, pp. 575–597, 2015.

[26] J. Cleland, J. M. Scobbie, C. Heyde, Z. Roxburgh, and A. A. Wrench, "Covert contrast and covert errors in persistent velar fronting," *Clinical Linguistics and Phonetics*, vol. 31, no. 1, pp. 35–55, 2017.

[27] B. Dodd, H. Zhu, S. Crosbie, A. Holm, and A. Ozanne, *Diagnostic evaluation of articulation and phonology (DEAP)*. Psychology Corporation, 2002.

[28] J. McCann and A. A. Wrench, "A new EPG protocol for assessing DDK accuracy scores in children: a Down's syndrome study," in *Proceedings of the 16th International Congress of the ICPhS*, 2007, pp. 1985–1988.

[29] K. Richmond, R. Clark, and S. Fitt, "On generating Combilex pronunciations via morphological analysis," in *INTERSPEECH*, 2010, pp. 1974–1977.

[30] K. Richmond, R. A. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *INTERSPEECH*, 2009, pp. 1295–1298.

[31] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2009.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[33] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ–distortion and discrete/continuous HMM's," *IEEE Transactions on speech and audio processing*, vol. 2, no. 3, pp. 456–459, 1994.

[34] H. Bredin, "pyannote. metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *INTERSPEECH*, 2017.

[35] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[36] J. Cleland, A. Wrench, S. Lloyd, and E. Sugden, "Ultrax2020: Ultrasound technology for optimising the treatment of speech disorders: Clinicians' resource manual," University of Strathclyde, Tech. Rep., 2018.